

## REMARKS

### The Section 112 Rejection

Claim 17 was rejected because the term “expectancy” was allegedly vague and indefinite under Section 112, second paragraph. Applicant has amended this language in claim 17, based on the specification at page 18, last paragraph. Thus, applicant respectfully requests withdrawal of the Section 112 rejection.

### The Patentability Rejections

Claims 14-15 and 19-22 were rejected under Section 102(b) based on Eisen et al., and claims 14-17 and 19-22 were rejected under Section 103(a) over Eisen in view of Altschul.

The present claims patentably distinguish from the cited references for the reasons set forth below.

Applicant will first review some of the context and embodiments of the invention as described in the specification.

As described in the “Overview” section on pages 9-10, methods and systems consistent with embodiments of the present invention enable the extraction of attributes from sequence strings and from information representing biopolymer materials. Sequence strings are useful to provide an abstract representation of a complex object. In the biosciences, sequence strings are used to represent biopolymer materials, which are macromolecules found within a living thing, such as proteins, nucleic acids, etc.

However, the sequence strings of biopolymer materials are difficult to analyze, especially in large numbers. Most people can only remember seven units of unrelated

information at a time, and sequence strings can be much longer. The need for analysis of a large number of biopolymer materials makes such analysis by a person impractical. Computer processing makes analysis easier, however the computer must be able to understand the sequence string. Accordingly, it is desirable to provide a computer with understandable attributes of biopolymer materials, whereby relationships among biopolymer sequence strings can be examined simultaneously.

As further described on pages 11-12, Fig. 2 shows a step of a computer process to visualize a data set representing a biopolymer material, e.g., a protein data set. A context vector is created for each of the biopolymer materials, e.g., proteins in the data set. The context vector is of a high dimension so as to represent many attributes of the material. The context vector can be used to visualize the data set to find related or unrelated attributes of the biopolymer materials.

With reference to Fig. 3, one method for creating context vectors for biopolymer materials uses sequence data. In this method, each biopolymer material in the data set is identified as a respective series of sequence letters. However, the series of sequence letters lack ascertainable attributes. Therefore, it is necessary to use special processing to determine the attributes of this sequence data.

This is one of the problems addressed by applicant's invention, namely that the sequence data is relatively undefined -- we don't have a measure or indication of what the sequence data is about. According to the invention, methods and systems are provided for comparing sequence data, for which we do not have a measure of distinguishing information, in order to generate comparison data. From the comparison

data we create high-dimensional context vectors, and from the context vectors we create a comparison matrix which can be used for visualizing the data set.

Applicant now refers to the language of claim 14 which is illustrated by an embodiment set forth in the specification at pages 18-22.

In the preamble, claim 14 recites a method for generating a high-dimensional vector for at least one of a plurality of biopolymer materials represented in a set of sequence data. One embodiment, described beginning on page 18, is a method for creating context vectors corresponding to biopolymer material using an indirect, relative and non-geometric indication of the structure or function of the biopolymer material. The embodiment refers to analysis of a protein as an example of a biopolymer material.

Claim 14 further recites the step of comparing sequence data regarding each biopolymer material to sequence data regarding each other biopolymer material to provide a respective comparison result. The embodiment on page 18 describes comparing each protein structure to a data set of proteins to determine each protein's similarity to each of the other proteins. One method, utilizing the Basic Local Alignment Search Tool (BLAST), provides a list of proteins from a data set ranked in order by expect values or probability scores. The score is usually expressed as an expectation that in comparing one sequence against a number of other sequences a match would have been found, wherein a very low probability score would say that there was very little expectation to have found a match by chance and thus the sequences must be related. The BLAST method is said to give a good picture of similarity and can be used to provide a non-geometric distance measure.

Claim 14 further recites the step of arranging the comparison results in a square matrix indexed by the plurality of biopolymer materials. An embodiment in the specification beginning on page 20 describes placing the expect scores from each comparison in a square matrix which represents the data set (illustrated in Fig. 7).

Claim 14 then recites the step of creating a high-dimensional context vector for one of the biopolymer materials based on a row or column of the square matrix. Continuing at page 20 of the specification, it is described how the rows or columns of the square matrix are used to create a context vector for each protein in the data set, by considering the entire row or column as a vector. For example, the i-th attribute in the vector is a comparison measurement between the i-th protein in the data set and the respective protein.

Claim 14 then recites the step of creating a comparison matrix based on the context vector to enable visualization of the sequence data of the respective biopolymer material. The specification at pages 21-22 describes how once the context vectors are provided, they can be projected into a two or three dimensional viewing area, identified as step 230 in Fig. 2. Furthermore, clustering may then be used to determine a centroid for a subset of proteins, and then the clusters and objects, e.g., proteins, can be projected onto a two- or three-dimensional viewing area. Applicant's related application Serial No. 09/408,716 further describes a method of visualizing context vectors based on sequence data.

In contrast to the method described in claim 14, Eisen starts with a series of gene expression data. This expression data constitutes a measure of what each gene is about. Therefore, Eisen is completely unrelated to the problem of the present invention

and the claimed solution. Eisen does not describe applicant's four steps for creating a comparison matrix because Eisen starts out with a data set which does provide a direct measure or indication of the behavior of the objects (in that case genes) -- namely their expression data. Thus, Eisen can proceed directly to clustering his set of gene expression data, wherein during clustering Eisen forms a distance matrix. Again, this process described by Eisen is something which could potentially occur after performing applicant's method, for example if applicant wished to perform a cluster analysis on the comparison results in applicant's comparison matrix. Again, however, Eisen fails to teach or suggest the four-step method of creating a comparison matrix as recited in claim 14.

Thus, in this case not only is applicant having to derive the indication, but it is a measure of the relatedness of the actual biopolymer string instead of an unrelated attribute like expression. Dependent claims 15-17 and 19 further describe the method of claim 14, with regard to: providing context vectors for each biopolymer material in the set (claim 15); utilization of the BLAST tool for the comparison step (claim 16); providing the comparison results based on an expectation of a relation (claim 17); and specifying the biopolymer material as a nucleic acid (claim 19).

Claim 20 defines an apparatus, including a memory having program instructions and at least one processor configured to execute the program instructions to perform the operations of a method; that recited method is the same as recited in claim 14.

Claim 21 defines an apparatus as a series of means for performing the steps according to the same method as recited in claim 14.

Claim 22 defines a computer-readable medium containing instructions for controlling a computer system to perform the same method as recited in claim 14.

Thus, each of the remaining independent claims 20-22 are patentable for the same reasons as claim 14.

In view of the foregoing amendments and remarks, applicant respectfully requests the reconsideration and reexamination of this application and the timely allowance of the pending claims.

Please grant any extensions of time required to enter this response and charge any additional required fees to our deposit account 06-0916.

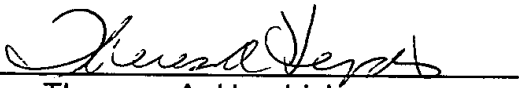
Respectfully submitted,

FINNEGAN, HENDERSON, FARABOW,  
GARRETT & DUNNER, L.L.P.

Dated:

*Jan 10, 2003*

By:

  
Therese A. Hendricks  
Reg. No. 30,389

## MARKED UP CLAIMS

14. A method for generating a high-dimensional vector [representation of an item of biopolymer material in a data set including] for at least one of a plurality [of items] of biopolymer [material,] materials represented in a set of sequence data, the method comprising:

comparing [information] sequence data regarding each biopolymer material [of the plurality to information] to sequence data regarding each other biopolymer material to provide a respective comparison result;

arranging the comparison results in a square matrix indexed by the plurality of [items of] biopolymer materials;

creating a high-dimensional context vector for [an item of] one of the biopolymer materials based on a row or column of the square matrix; and

creating a [distance] comparison matrix based on [the high-dimensional vector] the context vector to enable visualization of the sequence data of the respective biopolymer material.

15. The method according to claim [13] 14, wherein from each row or column of the square matrix, a respective high-dimensional context vector is created for each of the [items of] biopolymer materials based on the comparison results in the row or column.

16. (unamended) The method according to claim [13] 14, wherein the comparing uses a Basic Local Alignment Search Tool.

17. The method according to claim [13] 14, wherein the comparing provides comparison results based on an [expectancy] expectation of a relation.

19. The method according to claim [13] 14, wherein the biopolymer material is nucleic acid.

20. An apparatus for generating a high-dimensional vector [representation of an item of biopolymer material in a data set including] for at least one of a plurality [of items] of

biopolymer [material,] materials represented in a set of sequence data, the apparatus comprising:

at least one memory having program instructions, and

at least one processor configured to execute the program instructions to perform the operations of:

comparing [information] sequence data regarding each biopolymer material [of the plurality to information] to sequence data regarding each other biopolymer material to provide a respective comparison result;

arranging the comparison results in a square matrix indexed by the plurality of [items of] biopolymer materials;

creating a high-dimensional context vector for [an item of] one of the biopolymer materials based on a row or column of the square matrix; and

creating a [distance] comparison matrix based on [the high-dimensional vector] the context vector to enable visualization of the sequence data of the respective biopolymer material.

21. An apparatus for generating a high-dimensional vector [representation of an item of biopolymer material in a data set including] for at least one of a plurality [of items] of biopolymer [material,] materials represented in a set of sequence data, the apparatus comprising:

means for comparing [information] sequence data regarding each biopolymer material [of the plurality to information] to sequence data regarding each other biopolymer material to provide a respective comparison result;;

means for arranging the comparison results in a square matrix indexed by the plurality of [items of] biopolymer material;

means for creating a high-dimensional context vector for [an item of] one of the biopolymer materials based on a row or column of the square matrix; and

means for creating a [distance] comparison matrix based on [the high-dimensional vector] the context vector to enable visualization of the sequence data of the respective biopolymer material.



22. A computer-readable medium containing instructions for controlling a computer system to perform a method for generating a high-dimensional vector [representation of an item of biopolymer material in a data set including] for at least one of a plurality [of items] of biopolymer [material,] materials represented in a set of sequence data, the method comprising:

comparing [information] sequence data regarding each biopolymer material [of the plurality to information] to sequence data regarding each other biopolymer material to provide a respective comparison result;

arranging the comparison results in a square matrix indexed by the plurality of [items of] biopolymer materials;

creating a high-dimensional context vector for [an item of] one of the biopolymer materials based on a row or column of the square matrix; and

creating a [distance] comparison matrix based on [the high-dimensional vector] the context vector to enable visualization of the sequence data of the respective biopolymer material.

## BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate the implementations of the invention and together with the description, serve to explain the principles of the invention.

FIG. 1 is a diagram of an exemplary computing system with which the present invention may be implemented;

FIG. 2 is a flow chart of steps used to visualize a data set of biopolymer material consistent with the present invention;

FIG. 3 is a flow chart of steps of an implementation of a method for creating context vectors for sequence strings;

FIG. 4 is a flow chart of steps of an implementation of a method for creating context vectors for biopolymer ~~material~~<sup>materials</sup>; *utilizing predefined domains*

FIG. 5 is a flow chart of steps of another implementation of a method for creating context vectors for biopolymer ~~material~~; *materials, utilizing a geometric shape*

FIG. 6 is a flow chart of steps of another implementation of a method for creating context vectors for biopolymer ~~material~~; *and materials utilizing a non-geometric indication of the biopolymer material*

Fig. 7 is an illustration of a square matrix created in the method of Fig. 6.

## DETAILED DESCRIPTION OF THE INVENTION

Reference will now be made in detail to the construction and operation of an implementation of the present invention which is illustrated in the accompanying

high-dimensional space at a predetermined amount. Alternatively, the number of occurrences of an n-gram word in the protein will be used to increase a magnitude of the context vector along the corresponding axis of the word in the high-dimensional space in proportion to the number of occurrences. The absence of a word will result in a zero value for the corresponding axis of the word in the high-dimensional space. Thereby, a context vector for each protein in the data set is created.

The use of n-gram words to create context vectors can also be used for sequences representing things other than protein. For example, nucleotide sequences have an alphabet of only four letters (G, A, T, C). When the alphabet is reduced, the word length of the n-gram can be increased.

## 2. Predefined Domains

A second method for creating context vectors for biopolymer material uses predefined domains to define the attributes of the biopolymer material.

Generally speaking, proteins have evolved from a set of building blocks with each protein arising from a different combination of these building blocks. In proteins, building blocks are known as motifs, and represent structural or functional domains. All proteins are built from the same sets of motifs. Current research has identified approximately 5000 motifs so far.

In this method, predefined domains of interest in the protein data set, such as motifs, are selected (see Fig. 4, step 410). For example, a user could select a file including motif definitions from a public domain set, such as PROSITE (available at various locations including <http://www.expasy.ch/prosite/>), or could select any

if the descriptor is present or zero if absent. Like predefined domains, the value along each axis could alternatively be scaled by the number of occurrences for each descriptor.

#### 4. Non-Geometric Indication

Fig. 6 is a flow chart of another method for creating context vectors corresponding to biopolymer material using an indirect, relative, and non-geometric indication of the structure or function of the biopolymer material, rather than an indicator of the actual structure or portion thereof (by, e.g., using n-gram words, motifs, or surfaces).

A common method used for analysis of biopolymer material, e.g., protein, involves comparing each protein's structure to a data set of proteins to determine each protein's similarity to each of the other proteins. One conventional method, the Basic Local Alignment Search Tool (BLAST), provides a list of proteins from the data set rank ordered by expect values. Various entities provide BLAST algorithms including <http://www3.ncbi.nlm.nih.gov/BLAST/>. The BLAST method provides a probability score when comparing one sequence against another. The score is usually expressed as the 'expectation' that in comparing the test sequence against a number of other sequences a match would have been found, e.g., a probability score of  $1 \times 10^{-180}$  would say that there was very little expectation to have found the match by chance and thus the two sequences must be related. A score of close to one indicates that the match would have been expected by chance, i.e., the homology is very weak. To provide the probability score, BLAST uses a heuristic algorithm that seeks local as opposed to

onto the two- or three-dimensional viewing area. Previously discussed U.S. Patent Application Serial No. 09/408,716, entitled DATA PROCESSING, ANALYSIS, AND VISUALIZATION SYSTEM FOR USE WITH DISPARATE DATA TYPES, describes one method of visualizing context vectors based on sequence data.

D. Conclusion

While there has been illustrated and described what are at present considered to be a preferred implementation and method of the present invention, it will be understood by those skilled in the art that various changes and modifications may be made, and equivalents may be substituted for elements thereof without departing from the true scope of the invention.

Modifications may be made to adapt a particular element, technique, or implementation to the teachings of the present invention without departing from the spirit of the invention. For example, any living material, from organism to microbe, could be represented using the context vectors of the present invention. Further, the present invention is not limited to the biosciences, and any material or energy could also be represented.

Also, the foregoing description is based on a client-server architecture, but those skilled in the art will recognize that a peer-to-peer architecture may be used consistent with the invention. Moreover, although the described implementation includes software, the invention may be implemented as a combination of hardware and software or in hardware alone. Additionally, although aspects of the present invention are described